Statement of Research

Sudharshan Vazhkudai
Argonne National Laboratory
The University of Mississippi

My research interests span several areas of designing, building and measuring complex systems, such as operating systems, clusters, and massively distributed Grid systems. My research is often driven by the need for enhanced performance or by the need to facilitate efficient resource management of large-scale settings. In some cases, I have focused on improving performance and accuracy, for example, by enhancing the throughput of a cluster communication subsystem, reducing the prediction error in bulk Grid data transfer forecasts, or improving the performance of parallel downloads in Grids. In other cases, I have addressed the complexities of developing scalable scheduling or co-allocation architectures such as with the Globus Toolkit™ or building environments for load sharing in clusters.

My research largely revolves around the following two aspects:

- The recent proliferation of high-speed networks and increasing improvements in storage and computing technologies have enabled the collaboration of disparate, massively distributed communities spanning multiple administrative domains. Fuelling these trends is the ever-increasing application requirement for tremendous amounts of computing, storage, high transfer speeds, and stringent quality-of-service guarantees. Today, several applications (ranging from genomics to earthquake simulations to high-energy physics) rely heavily on the efficient use of aggregated resources (spanning several domains) provided by the Grid. I am interested in the various aspects of Grid-related research such as job/data management/scheduling, quality-of-service guarantees, information discovery and dissemination, replication, and incentives for collaboration.

- The ability to construct clusters of workstations from off-the-shelf components can be exploited to aggregate resources for use by the Grid. I am also interested in the efficient resource management and performance of local-area distributed systems.

Research Philosophy

My approach to research is both applied and experimental – testing ideas by building real, working systems, measuring performance under realistic conditions, and then analyzing to derive insights on the design. It is vitally important that research systems be deployed and tested out incrementally, to identify design flaws and scalability issues early in the design cycle.

I am a strong proponent of research leading to usable deliverables and prototypes, apart from publications. To this end, several pieces of my doctoral work have been made available as part of the Globus Toolkit™ and are used as a base for future work at institutions such as Argonne National Laboratory, the University of Southern California, INFN (Italy), Fermilab, CERN, and University College (London). Further, my work on distributed Linux has been made available for educational purposes to graduate students from several institutions.

I believe in open source ideologies and have contributed to premier projects such as Globus Toolkit™ and Linux. Further, codes to most of my projects have been made publicly available. In addition, I have made my research results and data available to other researchers. Researchers from the University of California at Santa Barbara, for example, recently presented their work at Supercomputing 2002 using my research data on bulk transfer predictions to validate their methods.

Bulk Data Transfer Forecasts and the Implications to Grid Scheduling

Increasingly, scientific discovery is driven by computationally intensive analyses of massive data collections stored in a distributed fashion. Technology trends (Moore's and Gilder's laws) indicate that the rates at which networks and storage elements double in capacity or halve in price are approximately 8 and 12 months, respectively. This promising recent trend is propelling several experiment groups to undertake projects never before foreseen – illustrious examples being the search for extraterrestrial intelligence (SETI@Home), studies on protein folding (Folding@Home), analysis of aggregated minute genetic mutations causing major evolutionary changes (Evolution@Home), high-energy physics experiments (CMS), studies of gravitational waves (LIGO), and mapping of a quarter of the visible sky (digital sky survey, SDSS), etc. The common denominator in all of these projects is the propensity to use resources as distributed data stores or data staging facilities, which has thrust to the forefront the need to efficiently manage data access in massively distributed communities.

Attempts to build solutions for such distributed data intensive science can be broadly classified under the *Data Grid* campaign – The Grid Physics Network (GriPhyN), Particle Physics Data Grid, and European Data Grid, being a few examples. Data Grids are essentially federations of high-end storage systems attempting to build resource management infrastructures by providing solutions for securely accessing remote resources, identifying and publishing information on the Grid fabric (compute, storage, network elements), resource discovery, job management, data movement, and the like, thus enabling and propelling distributed petascale science. Grids, in general, is stretching the frontiers of distributed computing and are often compared with and expected to mature and evolve similar to the Internet.

Data Grid environments typically replicate datasets across a geographically dispersed set of resources, often for fault tolerance or performance reasons. For instance, the GriPhyN project proposes to replicate the high-energy physics data (several petabytes) using a tiered architecture, wherein all the data (approximately 20 petabytes by 2006) is located at a single Tier 0 site; various (overlapping) subsets of this data are located at national Tier 1 sites, each with roughly one-tenth the capacity; smaller subsets are cached at smaller regional Tier 2 regional sites; and so on. As such Data Grids begin to be deployed, datasets are prone to be replicated further for performance and proximity reasons. For instance, a physicist interested in a particular dataset might cache it locally for future access and might even publish it for others.

My work addresses the issues involved in the efficient selection and access of replicated data in Data Grid environments. My work has been primarily in the context of the Globus Toolkit™ (delivering technologies for the GriPhyN project), building middleware that (1) selects replicas in highly replicated environments, enabling efficient scheduling of data transfer requests; (2) predicts transfer times of bulk wide-area data transfers using extensive statistical analysis; and (3) co-allocates bulk data transfer requests, enabling parallel downloads from mirrored sites.

Scalable Replica Selection: Since replicated sites may have varying performance characteristics (because of diverse storage system architectures, network connectivity features, or load characteristics), users may want to be able to determine the site from which particular data sets can be retrieved most efficiently, especially as data sets of interest tend to be large (1 MB – 1 GB). To address this *replica selection* problem I have designed a decentralized storage brokering strategy wherein every client that requires access to a replica performs the selection process rather than a central manager performing matches against clients and replicas. I have further analyzed the prototypes in a wide-area testbed [1, 2].

Predictions of Grid Data Transfers: The selection strategy requires information about the capabilities and performance characteristics of storage systems (replica locations). Thus, of significant interest is the speed of a storage system, or, rather, the time a storage system takes to deliver a replica. Intelligent replica selection can be achieved by having replica locations expose performance information about past data transfers. This information can, in theory, provide a reasonable approximation of the end-to-end throughput for a particular transfer. It can then be used to predict future behavior between the sites involved. I have applied these techniques to the GridFTP (de facto data movement tool in Grids), part of the Globus Toolkit™.

Forecasting data transfer times involved the development of a series of univariate [3, 6] and multivariate [4, 5, 6] predictors deriving predictions from past history of GridFTP transfers in isolation and a combination of several data sources respectively. Univariate predictors include mathematical models such as mean-based, median-based and autoregressive tools and achieve acceptable levels of accuracy given their ease of implementation. Univariate models fail to account for system or network variations and the sporadic nature of Grid data transfers. Multivariate predictors address such factors by combining end-to-end application throughput observations with network and disk load variations, capturing whole-system performance and variations in load patterns, respectively. These predictions thus characterize the effect of load variations of several shared devices (network and disk) on file transfer times. Multivariate regression-based forecasting tools serve as a means both of improving accuracy and as a means of handling Data Grid variations (availability of data, sporadic nature of transfers, etc.). I have performed extensive performance analysis in Data Grid testbeds and observed performance predictions within 15% error, which is quite promising for a pragmatic system.

Co-Allocation of Grid Data Transfers: It is quite possible for more than a single site (hosting the replica) to offer excellent performance ratios. Or, in some cases, a slow server might be servicing a fast client, where the *out-bandwidth* is much lesser than the *in-bandwidth*. I have developed several co-allocation techniques that brokers can use to fetch partial copies of the replica in parallel. The techniques include (1) history-based discrimination of flows wherein the data size per flow is commensurate with the predictive merit of that replica location and (2) load-balancing techniques wherein the load between co-allocated flows is dynamically altered so that faster servers deliver larger chunks of data. More specifically, I have developed conservative and aggressive load-balancing mechanisms that address limitations of previous approaches. These mechanisms are part of the Data Grid middleware that can automatically download different parts of a file from multiple sources.

Fabric Support for Grids

Grid systems construct complex overlays of metaschedulers, information directory services, security mechanisms, and so forth, atop local resource management entities. For instance, Grid schedulers eventually delegate jobs to local schedulers; Grid security mechanisms will need to work hand-in-hand with local authentication. Essentially Grids harness the collective potential of local resource management systems. To this end, I have led the design and implementation of a distributed Linux with kernel-level modifications for enhanced performance [9]. I have designed, implemented, and analyzed a cluster communication subsystem with capabilities for multiplexing packets across multiple network channels, short-circuiting the overheads of TCP stack, memory copies, and user-kernel switches [7]. Higher-level services such as schedulers, distributed file systems, and local information services were constructed on this communication architecture exploiting direct network mapped virtual circuits. Distributed file access performance was considerably improved as a result of the communication infrastructure, prefetching, and an assumed mounts strategy [8]. I have also developed the architectural support required for remote job execution and the communication of input, error, and output back to the host machine (for instrumented applications) [9]. Such a system can in principle, be plugged into the Grid for use by other entities. My work on this system led to the creation of a fully functional distributed Linux testbed, which was used as a base for research by several students for their projects and theses.

Future Research and Plans

My future interests are directed toward understanding how to build large-scale systems with complex resource-sharing agreements. My long-term goal is to evolve towards an incentive-based resource-sharing system, wherein resource usage can be reserved, accounted, metered, and charged [10, 11]. Instead of developing such solutions immediately, however, I would like to adopt a bottom-up approach of prototyping, analyzing, and understanding various resource-sharing paradigms, and resource management systems, and evolving in the process toward sophisticated structures. For instance, my previous work of prediction-based resource access or co-allocated access can be viewed in this light. To further my understanding, I would like to delve into the use of quality of service for resource usage and the use of contracts and guarantees, reservations, and incentives in Grid environments.

In the near term, I am interested in exploring the ramifications of my previous work on several aspects of Grid research in general. Strategies for the efficient, scalable replication of bulk scientific data in experiment communities are essential for distributed data intensive science. To this end, I am interested in exploring peer-to-peer systems. Data, once replicated, needs to be scheduled efficiently alongside computation. Scheduling for massively distributed resource management systems pose new questions regarding scalability, co-allocations of compute and storage resources, coupling of data and computation that are disparately located, etc. My work on replica selection, addressed a narrow aspect of this problem. I am thus interested in exploring several issues regarding efficient construction of middleware for scheduling and resource co-allocation in Grids. Storage system guarantees for space reservations, pinning of data, and storage quality of service are becoming important with stringent application requirements. Either fetching data to run the computation locally or staging data at an intermediate location requires different types of guarantees from underlying storage management systems. I am interested in exploring mechanisms for dynamic quota management, pinning of data for future use, storage quality of service, advance reservations, in the context of distributed data intensive science. I am further interested in enabling the efficient access bulk data and tuning data movement protocols to obtain high performance for wide-area environments.

In the long term, I am interested in constructing massively distributed systems with different levels of motivations and incentives. Let us assume for a moment the existence of reasonable motivation for storage resources to host replicas/datasets. These are essentially experiment groups coming together to share datasets of interest. Initially, such groups could be scientists themselves hosting data with the hope that they can get access to data not in their possession – Condor-like motivation (contribute your resource to the pool if you wish to gain from the pool). Or, incentives could take the form of recognition in the community – many users desiring to access datasets from a particular host, denoting confidence in the quality and integrity of the data, and so on. With the existence of some motivation, what would a storage location or a storage service provider (SSP) do in order to sustain a user base? One can draw analogies from e-commerce Websites such as Amazon.com – (1) provide quality of service (2) study temporal evolutions of request access patterns to understand the kinds of data users request, and (3) develop mechanisms to acquire and host data not available. SSPs could themselves enter into agreements or contracts to exchange datasets they currently do not possess and are often requested. This can eventually evolve into a sophisticated *overlay network of storage service providers* with storage outsourcing and data trading capabilities. One can draw numerous analogies from the peer-to-peer world in terms of constructing and managing overlays. Research is needed that involves mechanisms and protocols to establish contracts, duration, terms of agreement, detecting violations, accountability, jurisdiction, and so forth. We could eventually evolve to *bartering* style resource sharing – users entering into contracts or agreements with resources in return for some quality of service. Service agreements can take the form *I wish to access the replica at this time in the future. Can you pin the replica for me?* or *I wish to stage data and require storage up to 20G for a duration of 20 hours*.

The next step in such collaborative e-Science could be the involvement of commercial service providers. Gradually, we can envision a *SETI-style storage* wherein ordinary users contribute their resources for storing scientific data (widely replicated of course) in return for some incentive. Thus, distributed data-intensive science, powered by the Grid, can progressively move from specialized scientific communities to the induction of commercial service providers and eventually to the masses with different levels of motivation and sharing dynamics similar to the evolution of the Internet. We can easily imagine incentives for accessing all resource types, such as processors, sensors, displays, networks, and job submissions, and several market type formulations.

To achieve this level of complexity in Grid systems, one needs a fairly deep knowledge of the dynamics of sharing in large-scale environments. Such systems are prone to constant transformation, with ever-evolving user requirements and sharing dynamics. I plan to investigate various issues involved in this domain, focusing on Grid computing areas such as distributed scheduling, contracts, quality of service, data management, incentives, and market formulations.

References

[1] S. Vazhkudai, S. Tuecke, I. Foster, "Replica Selection in the Globus Data Grid", *Proceedings of the IEEE International Conference on Cluster Computing and the Grid (CCGRID 2001),* pp. 106-113, Brisbane, Australia, May 2001.

[2] S. Vazhkudai, S. Tuecke "A Storage Broker for the Globus Environment - A ClassAd Based Implementation", *Poster in Supercomputing 2000,* Dallas, Texas, Nov 2000.

[3] S. Vazhkudai, J. Schopf, I. Foster, "Predicting the Performance of Wide-Area Data Transfers", *Proceedings of the 16th International Parallel and Distributed Processing Symposium (IPDPS 2002),* Fort Lauderdale, Florida, April 2002.

[4] S. Vazhkudai, J. Schopf, "Predicting Sporadic Grid Data Transfers", *Proceedings of the 11th IEEE International Symposium on High Performance Distributed Computing (HPDC-11),* pp. 188-196, Edinburgh, Scotland, July 2002.

[5] S. Vazhkudai, J. Schopf, "Using Disk Throughput data in Predictions of End-to-End Grid Transfers", *To appear in 3rd International Workshop on Grid Computing (GRID 2002),* Baltimore, Maryland, November 2002.

[6] S. Vazhkudai, J. Schopf, "Using Regression Techniques to Predict Large Data Transfers", *Submitted to the Journal of High Performance Computing Applications - Special Issue on Grid Computing: Infrastructure and Applications.*

[7] S. Vazhkudai, P.T. Maginnis, "A High Performance Communication Subsystem for PODOS", *Proceedings of the First IEEE International Conference on Cluster Computing*, pp. 81-91, Melbourne, Australia, December 1999.

[8] S. Vazhkudai, P.T. Maginnis, "The PODOS File System - Exploiting the High-Speed Communication Subsystem", *Proceedings of the IEEE International Workshop on Cluster Computing Technologies, Environments, & Applications*, pp. 453-459, Las Vegas, Nevada, June 2000.

[9] S. Vazhkudai, J.M. Syed, P.T. Maginnis, "PODOS - The Design and Implementation of a Performance Oriented Linux Cluster", *Journal of Future Generation Computer Systems - Special Issue on Cluster Computing*, Volume 18, Issue 3, pp. 335-352, January 2002.

[10] R. Buyya, S. Vazhkudai, "Compute Power Market: Towards a Market-Oriented Grid", *Proceedings of the IEEE Session on Global Computing on Personal Devices,* pp. 574-581, Brisbane, Australia, May 2001.

[11] S. Vazhkudai, G.V. Laszewski, "A Greedy Grid - The Grid Economic Engine Directive", *Proceedings of the IEEE Workshop on Internet Computing and E-Commerce,* San Francisco, California, April 2001.